UN MODELO LINEAL GENERALIZADO SEMIPARAMETRICO PARA ANALISIS DE DURACION CON CENSURA

Jesus Orbe[1]

Departamento de Econometría y Estadística (E.A. III)
Universidad del País Vasco/Euskal Herriko Unibertsitatea
Avenida Lehendakari Aguirre, 83
E-48015 Bilbao
E-mail: jo@alcib.bs.ehu.es

[1]Deseo expresar mi agradecimiento a Eva Ferreira y Vicente Núñez por sus comentarios y sugerencias, asimismo señalar que este trabajo ha sido financiado por los proyectos de investigación UPV 038.321-HA129/99 de la Universidad del País Vasco/Euskal Herriko Unibertsitatea y PB98-0149 de la Dirección General de Enseñanza Superior e Investigación Científica del Ministerio Español de Educación y Cultura.

Resumen

Aitkin y Clayton (1980) proponen el análisis de modelos de duración mediante modelos lineales generalizados. En este trabajo extendemos esta metodología permitiendo que el efecto de alguna de las variables explicativas pueda no ser especificado. Así, el modelo propuesto es un modelo lineal generalizado semiparamétrico, con una componente paramétrica donde se especifica la forma funcional concreta del efecto de las variables explicativas sobre la duración, y una componente no paramétrica donde recogemos el efecto de una variable explicativa sin asumir forma funcional alguna. Desarrollaremos el proceso de estimación así como un procedimiento bootstrap para realizar inferencia. Como aplicación, analizaremos con la metodología propuesta el tiempo de supervivencia para una muestra de pacientes diagnosticados de SIDA.

Palabras clave: Modelos de duración, censura, bootstrap, supervivencia, estimación semiparamétrica

Clasificación AMS: 62J12, 62N05, 62G09

1 Introducción

Proponemos una extensión al trabajo presentado por Aitkin y Clayton (1980), quienes presentan la posibilidad de estimar una serie de modelos de duración paramétricos utilizando un modelo lineal generalizado para la variable indicador de censura, cuya función de verosimilitud es proporcional a la correspondiente del modelo de duración original. Tomando como base esta idea, extendemos la metodología de estos autores a un conjunto de situaciones más general. Esta extensión permite modelizar aquellas situaciones en las que la forma funcional del efecto de alguna de las variables explicativas sobre la variable de interés es desconocida o simplemente la especificación de una determinada forma funcional nos parece restrictiva. Esta flexibilización se puede realizar de un modo bastante natural basándonos en un modelo lineal generalizado. Así, vamos a extender el trabajo de Aitkin y Clayton a un contexto semiparamétrico. Como ilustración aplicamos la metodología para analizar el efecto de ciertas variables sobre el tiempo de supervivencia, desde el momento del diagnóstico, en una muestra de enfermos diagnosticados de SIDA.

Comenzamos con un breve repaso de los modelos lineales generalizados y de los modelos más comunes en análisis de duración.

1.1 Modelos Lineales Generalizados

Los Modelos Lineales Generalizados (MLG) son introducidos por primera vez en el trabajo de Nelder y Wedderburn (1972). Estos modelos pueden considerarse como una generalización del modelo lineal clásico:

$$Z = X\beta + \epsilon,$$

donde suponemos que los errores ϵ son independientes y tienen una distribución normal de media cero ($E(Z) = \mu = X\beta$), y varianza constante. La matriz X es una matriz $(n \times p)$ recogiendo las variables explicativas, $\beta^T = (\beta_1, \ldots, \beta_p)$ es el vector de coeficientes asociados a esos regresores y Z es la variable a explicar.

Podemos estructur este modelo clásico en tres partes:

- Componente aleatoria: formada por el vector que recoge los valores que toma la variable a explicar Z, con $Z \in N(\mu = X\beta; \sigma^2)$, y siendo σ^2 constante.
- Componente sistemática: donde tenemos las variables explicativas de X formando lo que se conoce como predictor lineal $\eta = X\beta$.
- Función de enlace: función que nos enlaza la media μ de la variable a explicar con la parte sistemática, proporcionando la linealidad en las variables explicativas. $g(\mu) = \eta = X\beta$. Para el modelo lineal clásico la función de enlace es la función identidad.

Este modelo puede utilizarse para analizar una gran variedad de situaciones, pero en algunos contextos resulta poco apropiado por diferentes motivos: la variable respuesta toma valores únicamente dentro de un intervalo concreto, hay relaciones no lineales entre la media de la variable a explicar y las variables explicativas, o bien la variable respuesta no tiene varianza constante, etc.

Los MLG permiten incorporar este tipo de situaciones de forma que, mediante ciertas transformaciones adecuadas, se recupera la linealidad del modelo y, por tanto, las ventajas de los modelos lineales.

Así, el MLG puede considerarse como una generalización del modelo clásico, extendiéndolo a través de dos vías:

(i) vía función de enlace, permitiendo otra función de enlace monótona diferenciable distinta a la función identidad[2].

(ii) vía componente aleatoria: generalizando la estructura de normalidad o relajando el supuesto de varianza constante.

De esta forma, los MLG abarcan como casos particulares los modelos de regresión lineal, los modelos de análisis de varianza, modelos logit y probit para variables de respuesta dicotómicas, modelos log-lineales y multinomiales para variables de respuesta multinomial, además de algunos modelos habitualmente utilizados en el análisis de supervivencia o de datos de duración.

Todos ellos comparten una serie de propiedades tales como la linealidad y, además, los parámetros de interés pueden ser estimados utilizando el mismo método: maximizando la función de verosimilitud o, equivalentemente, realizando mínimos cuadrados ponderados iterativos[3]. McCullagh y Nelder (1983) y Fahrmeir y Tutz (1994) estudian en detalle diferentes modelos lineales generalizados.

1.2 Modelos de Duración

Como acabamos de indicar, los MLG pueden utilizarse para estimar modelos de duración. Si nos centramos en modelos de duración para poblaciones heterogéneas, es decir, en modelos de regresión, y realizamos un repaso a la abundante literatura en el tema, básicamente nos encontramos con dos grandes corrientes de modelos: los modelos de duración acelerada (ver, por ejemplo, Kalbfleish y Prentice, 1980) y los modelos de función de riesgo proporcional propuestos por Cox (1972).

Sea $T_1, \ldots, T_n$ una muestra aleatoria simple para la variable duración, la cual no es observada en su totalidad debido a la existencia de censura[4] y en su lugar observaremos

$$Y_i = \min(T_i, C_i), \qquad \delta_i = \begin{cases} 1; & \text{si } T_i \leq C_i \\ 0; & \text{si } T_i > C_i \end{cases},$$

donde $C_1, \ldots, C_n$ son los valores que toma la variable censura C, la cual suponemos independiente de la variable duración T. Además, δ_i es la variable indicador de censura, tomando valor 0 si la observación correspondiente está censurada o valor 1 si no lo está[5].

[2]Para ver las funciones de enlace más utilizadas, véase, por ejemplo, McCullagh y Nelder (1983), pág. 31.

[3]Es decir, utilizando el algoritmo conocido como Iterated Reweighted Least Squares (IRLS).

[4]Situación habitual cuando se analiza esta clase de datos. Si consideramos el caso más habitual, censura por la derecha, tenemos una observación censurada cuando su tiempo de fallo no ha sido observado al finalizar el estudio (para un detallado análisis de los distintos tipos de censura, vease, por ejemplo, Lawless, 1982).

[5]Con esta especificación estamos suponiendo el caso de censura aleatoria, el habitualmente utilizado. Además, añadiremos que este tipo de censura engloba a otros tipos de censura más restrictivos.

Una gran clase de modelos de duración tratan de modelizar la función de riesgo, $\lambda(t)$, definida como la probabilidad de fallar en el momento t, dado que se ha sobrevivido hasta ese momento. Es decir,

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}$$

El modelo de función de riesgo proporcional establece la siguiente modelización para la función de riesgo

$$\lambda(t, x) = \lambda_0(t) h(x, \beta),$$

donde $\lambda_0(t)$ es conocida como la función de riesgo básica y la especificación habitual para $h(x, \beta)$ es la función exponencial[6]. El procedimiento habitual de estimación para este modelo consiste en maximizar la función de verosimiltud parcial (Cox, 1975) sin suponer una forma funcional concreta para la función de riesgo básica o, lo que es lo mismo, sin suponer una distribución para la variable duración.

Si consideramos ahora la otra corriente importante de modelos de duración, los modelos de duración acelerada, tenemos la siguiente especificación para la función de riesgo

$$\lambda(t, x) = \lambda_0(t \cdot h(x, \beta)) h(x, \beta),$$

donde, para el caso habitual de $h(x, \beta) = e^{-x^T \beta}$, tenemos que el modelo puede reescribirse en términos log-lineales, donde tenemos una relación directa y lineal de las variables explicativas sobre el logaritmo de la variable T.

El procedimiento habitual de estimación en esta clase de modelos consiste en suponer una distribución para la variable duración y maximizar la función de verosimilitud

$$L = \prod_{i=1}^{n} f(y_i, x_i)^{\delta_i} S(y_i, x_i)^{1-\delta_i}, \tag{1}$$

donde la contribución de las observaciones no censuradas viene dada por la función de densidad, $f(t)$, y la de las censuradas por la función de supervivencia $S(t) = P(T \geq t)$.

En la práctica, el modelo más utilizado es el primero y la razón fundamental es la posibilidad de estimar los parámetros de interés sin suponer una distribución para la variable duración. Los modelos de duración paramétricos, como los modelos de regresión exponencial, Weibull, Gamma, Log-Normal o log-logístico, son casos particulares de la segunda clase de modelos, siendo únicamente los modelos de regresión Weibull y exponencial casos particulares de ambos grupos de modelos.

El resto del trabajo se organiza de la siguiente forma. En la Sección 2 presentamos la relación entre los MLG y algunos modelos de duración. En la Sección 3 mostramos una aplicación utilizando la metodología tradicional en análisis de duración y utilizando un MLG. Posteriormente, y motivándolo en la aplicación anterior, en la Sección 4, presentamos una extensión al trabajo de Aitkin y Clayton (1980) desarrollando el proceso de estimación de este nuevo modelo. En la Sección 5 proponemos un nuevo procedimiento bootstrap para realizar inferencia en el modelo propuesto. Finalmente, la Sección 6 presenta los resultados y conclusiones más importantes.

[6]La razón para utilizar esta especificación se basa en que garantiza la no negatividad de la función de riesgo sin imponer restricciones sobre los coeficientes β.

2 Conexión entre Modelos de Duración y Modelos Lineales Generalizados

Describiremos los pasos relevantes en la metodología de Aitkin y Clayton, para entender posteriormente la generalización propuesta. Para ello, consideraremos un conjunto de datos en los que la variable respuesta a analizar recoge la duración o el tiempo de supervivencia de un elemento poblacional. Supongamos que esa variable duración puede ser explicada con un modelo que pertenece a la clase de modelos de duración con función de riesgo proporcional

$$\lambda(t;x) = \lambda_o(t)\exp(x^T\beta),$$

donde $\eta = x^T\beta$ es el predictor lineal. La estimación de los parámetros del modelo puede realizarse maximizando la función de verosimilitud (1).

Utilizando las relaciones existentes entre las funciones de supervivencia, riesgo, riesgo acumulado ($\Lambda(t) = \int_0^t \lambda(t)dt$) y de densidad, obtenemos las siguientes expresiones:

$$S(t,x) = \exp(-\Lambda_0(t)e^{\eta})$$

$$f(t,x) = \lambda_0(t)\exp(\eta - \Lambda_0(t)e^{\eta}),$$

donde $\Lambda_0(t) = \int_0^t \lambda_0(t)dt$ es la función de riesgo acumulado básica.

Sustituyendo las expresiones anteriores en (1), tomando logaritmos y reordenando términos obtenemos,

$$\ln L = \sum_{i=1}^{n} \delta_i \left[\ln \Lambda_0(y_i) + \eta_i\right] - \Lambda_0(y_i)e^{\eta_i} + \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right). \tag{2}$$

Tomando $\mu_i = \Lambda_0(y_i)e^{n_i}$ tenemos que,

$$\ln L = \underbrace{\sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i)}_{(a)} + \underbrace{\sum_{i=1}^{n} \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)}_{(b)}. \tag{3}$$

Se puede verificar que el sumando (a) de la expresión anterior es proporcional al logaritmo de la función de verosimilitud correspondiente a una muestra de n variables aleatorias independientes δ_i con distribución de Poisson[7] de media μ_i. Por otra parte, el término (b) no depende de los parámetros β, sólo depende de la función de riesgo básica, la cual puede depender de parámetros de la distribución.

De esta forma, siguiendo el trabajo de Aitkin y Clayton (1980), y dada la función de riesgo acumulado básica $\Lambda_0(t)$, podemos estimar los coeficientes β del modelo tratando a la variable indicadora de censura δ_i como una variable aleatoria con distribución de Poisson de media $\mu_i = \Lambda_0(y_i)e^{\eta_i}$. Es decir, podemos construir un modelo log-lineal de Poisson "auxiliar" al modelo de duración, tal que

[7] $\sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i) - \sum_{i=1}^{n} \ln \delta_i!$

$$\ln(\mu_i) = \ln \Lambda_0(y_i) + x_i^T \beta,$$

con una función de verosimilitud proporcional al término (a) de la expresión (3).

Tenemos así una componente aleatoria con una distribución de Poisson, y una función de enlace logarítmica que nos relaciona a la media con el predictor lineal. Además en este modelo tenemos un término adicional, $\ln \Lambda_0(y_i)$, denominado "offset". El modelo log-lineal de Poisson, como caso particular de los MLG, puede ser maximizado utilizando el procedimiento de estimación habitual en éstos maximizando la función de verosimilitud o, equivalentemente, aplicando mínimos cuadrados ponderados iterativos (IRLS), cuyo algoritmo consta de los siguientes pasos:

Paso 1: Construir la variable Z^*, como la expansión de Taylor de primer orden, alrededor de la media μ, para la función enlace g aplicada sobre la variable respuesta Z.

$$g(z) = \underbrace{g(\mu) + (z - \mu)g'(\mu)}_{z^*} + \ldots,$$

$$z^* = \hat{\eta} + (z - \hat{\mu}) \left(\frac{d\eta}{d\mu} \right) \Bigg|_{\mu=\hat{\mu}}, \quad \text{donde} \quad \eta = g(\mu).$$

Paso 2: Calcular las ponderaciones,

$$W^{-1} = \left(\frac{d\eta}{d\mu} \right)^2 \Bigg|_{\mu=\hat{\mu}} V(\hat{\mu}),$$

donde $V(\hat{\mu})$ es la función de varianza de Z evaluada en $\hat{\mu}$. Por tanto, las ponderaciones coinciden con la varianza de Z^*.

Paso 3: Regresar de forma iterativa Z^* sobre las variables explicativas en X utilizando las ponderaciones calculadas en el paso 2. Es decir,

$$\hat{\beta} = (X^T W X)^{-1} X^T W Z^*$$

El proceso de estimación del modelo es menos directo si el término "offset" contiene parámetros desconocidos pertenecientes a la función de distribución de la variable duración.

En Aitkin y Clayton (1980) se describe en detalle el proceso de estimación mediante MLG para los modelos de duración con distribución exponencial, Weibull y valor extremo. Por otra parte, Whitehead (1980) propone un procedimiento de estimación análogo a los anteriores pero para aquellos casos en que desconocemos la forma funcional de la función de riesgo básica; es decir, propone la estimación mediante MLG para modelos de función de riesgo proporcional de Cox (1972).

3 Aplicación

3.1 Datos

Para ilustrar la metodología descrita hemos aplicado ésta para analizar el tiempo de supervivencia desde el momento del diagnóstico, en una muestra compuesta por 461 enfermos

diagnosticados de SIDA desde 1984 hasta el comienzo de 1991, residentes en las comunidades autónomas del País Vasco y Navarra. Utilizando la fecha de diagnóstico y la fecha de fallecimiento, o en el caso de las observaciones censuradas, la fecha final del seguimiento (diciembre de 1992), obtenemos la variable de interés, la duración o tiempo de supervivencia desde el momento del diagnóstico medida en número de trimestres. A diferencia de la mayoría de los trabajos realizados en este área, los cuales se han interesado en estudiar la duración del periodo de incubación, nosotros nos hemos centrado en el estudio de la duración de la última etapa. En el desarrollo del virus VIH tenemos tres etapas. La primera de ellas la conocida como fase "pre-anticuerpos" es la más corta con una duración de varios meses (aproximadamente el 50% de los enfermos genera anticuerpos antes de los dos meses después de la infección). Esta etapa va desde el momento en que se produce la infección hasta el desarrollo de los anticuerpos o punto de seroconversión, y es el periodo de tiempo donde al enfermo se clasifica como seronegativo. La segunda etapa, etapa de incubación, es la más larga de las tres (aproximadamente la mitad de los infectados desarrollaban la enfermedad antes de los 10 años). Este periodo parte desde el momento de la seroconversión hasta el diagnóstico de SIDA. Durante esta etapa el individuo es clasificado como seropositivo. Y por último, la tercera etapa, que recoge el tiempo de supervivencia desde el diagnóstico del SIDA. El comienzo de esta etapa tiene lugar en el momento en que el individuo desarrolla alguna enfermedad clasificada dentro de las enfermedades relacionadas con el SIDA.

Para ayudar a describir esta variable disponemos de una serie de variables que nos recogen ciertas características de los enfermos. Así, la variable **Edad** recoge la edad del enfermo en el momento del diagnóstico. **Sexo** es una variable ficticia, que toma valor 1 si el enfermo es varón y valor 0 si es mujer. Tenemos información sobre la enfermedad con la cual se le diagnostica el SIDA. Así, la variable **Enfer1** toma valor 1 si la enfermedad de diagnóstico es una infección oportunista, **Enfer2**, si es un linfoma o un sarcoma de Kaposi y **Enfer3** si es debido a una encefalopatía VIH o al síndrome de "agotamiento" VIH. Además, tenemos información sobre la vía de transmisión de la enfermedad: la variable indicador **Sexual** toma valor 1 si la vía de transmisión es sexual, **Drogas** toma valor 1 si la infección se produce por consumo de drogas, **Sanguínea** toma valor 1 cuando el enfermo es infectado por transmisión sanguínea, **Madre-hijo** toma valor 1 si la transmisión se produce de la madre al hijo, y **Otras** cuando se desconoce la vía de transmisión. Por último, la variable **Periodo** es una variable indicador que toma valor 1 cuando la fecha del diagnóstico es posterior a 1987. El motivo de introducir esta variable ficticia es estudiar el posible efecto de la introducción, a mediados de 1987, del fármaco Zidovudine (también conocido como AZT) sobre la supervivencia del enfermo.

3.2 Análisis de duración tradicional

Comenzamos el estudio del tiempo de supervivencia para los enfermos realizando un análisis preliminar no paramétrico. Es decir, no suponemos una distribución para la variable duración. Además, como primer paso, consideramos que la población de enfermos es homogénea; es decir, no introducimos el efecto que pudieran tener las variables explicativas sobre la duración. El objetivo de este primer análisis es realizar un análisis univariante de la duración y tratar de determinar la posible distribución de ésta. Por tanto, comenzamos estimando la

función de supervivencia mediante el estimador propuesto por Kaplan y Meier (1958) y la presentamos en la Tabla 1 junto con la estimación de las desviaciones típicas (Greenwood, 1926) y el correspondiente intervalo de confianza al 95%.

Tabla 1: Estimación Kaplan-Meier para $S(t)$[8]

Duración	n_j	d_j	$\hat{S}(t)$	$\hat{Des}(\hat{S}(t))$	lim. inf.	lim. sup.
0.5	461	106	0.770	0.0196	0.7335	0.8095
1.5	355	87	0.581	0.0229	0.5380	0.6282
2.5	268	40	0.495	0.0232	0.4510	0.5424
3.5	228	36	0.416	0.0229	0.3738	0.4640
4.5	192	39	0.332	0.0219	0.2916	0.3778
5.5	153	22	0.284	0.0210	0.2458	0.3285
6.5	131	22	0.236	0.0197	0.2007	0.2786
7.5	109	16	0.202	0.0186	0.1682	0.2419
8.5	93	17	0.165	0.0172	0.1342	0.2025
9.5	72	17	0.126	0.0155	0.0988	0.1605
10.5	53	13	0.095	0.0139	0.0713	0.1266
11.5	38	8	0.075	0.0126	0.0539	0.1044
12.5	30	4	0.065	0.0119	0.0454	0.0931
13.5	26	5	0.052	0.0108	0.0350	0.0788
14.5	20	3	0.045	0.0101	0.0286	0.0697
15.5	16	2	0.039	0.0096	0.0241	0.0633
16.5	12	3	0.029	0.0087	0.0164	0.0524
17.5	8	2	0.022	0.0079	0.0108	0.0445
18.5	6	1	0.018	0.0074	0.0083	0.0404
19.5	5	2	0.011	0.0059	0.0038	0.0319
21.5	3	1	0.007	0.0049	0.0019	0.0278
22.5	2	1	0.003	0.0035	0.0005	0.0251

Una vez estimada la función de supervivencia de la variable duración, ésta puede ser utilizada, mediante procedimientos gráficos, para tratar de determinar la posible distribución a la que se ajustan los datos.

Por ejemplo, podríamos llevar a cabo un contraste gráfico para ver si la duración tiene una distribución exponencial o Weibull (dos de las distribuciones más importantes y más utilizadas en el análisis de duración paramétrico). La función de supervivencia para una variable con distribución Weibull viene dada por:

$$S(t) = \exp[-(\lambda t)^p], \qquad \lambda > 0, \qquad p > 0, \qquad t > 0.$$

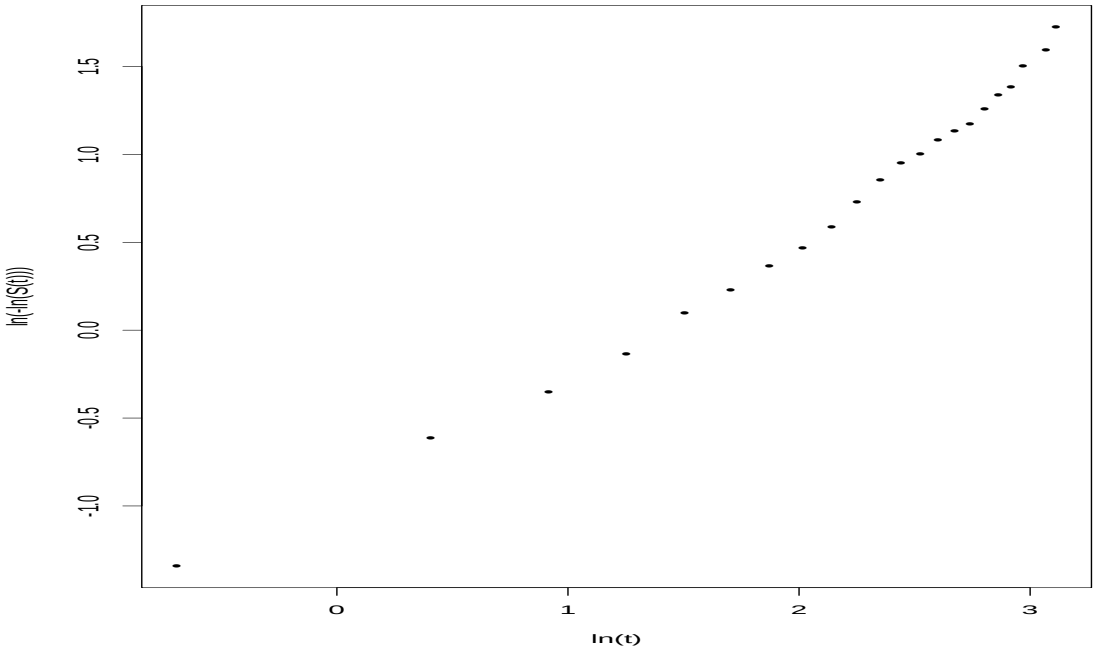
Así, si representamos $\ln[-\ln \hat{S}(t)]$ frente al logaritmo de las duraciones[9], $\ln(t)$, (Figura 1),

[8] n_j recoge el número de individuos en riesgo justo antes de t_j y d_j el número de individuos con una duración igual a t_j.

[9] Sumamos un valor de 0.5 a todas las duraciones para no tener problemas al tomar logaritmos.

y los datos se ajustan a una distribución de Weibull, deberíamos obtener aproximadamente una línea recta.

Figura 1: Contraste gráfico para una distribución Weibull



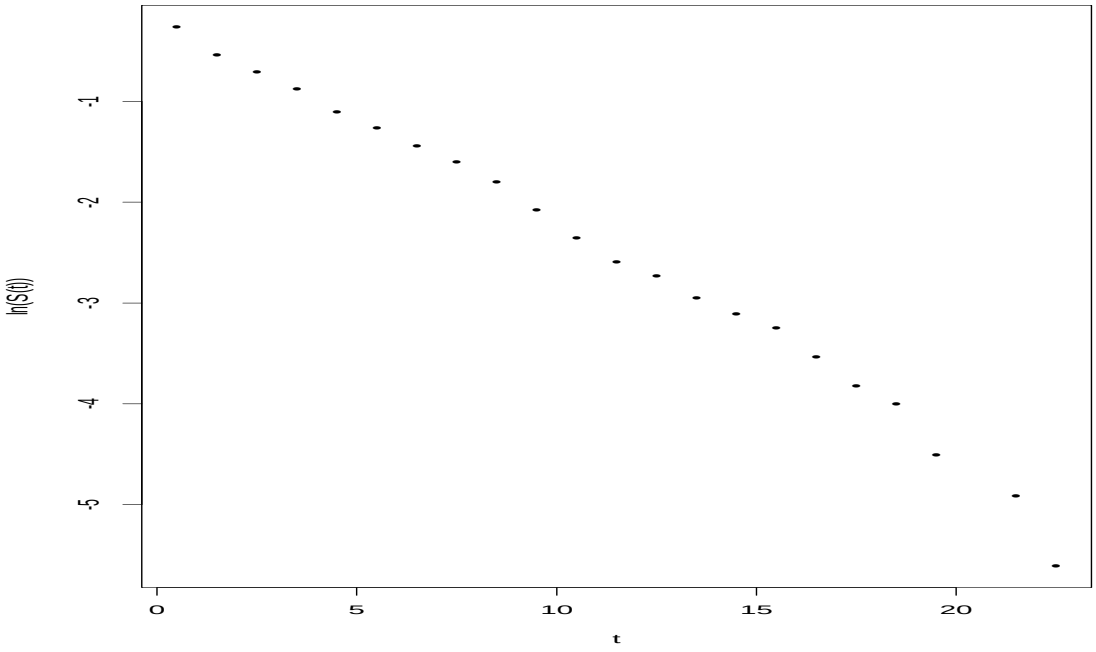
La Figura 1 indica que ésta distribución es adecuada para los datos.

Además, si observamos la pendiente del gráfico, tendremos una idea aproximada del valor para el parámetro de forma p. Por tanto, parece adecuado suponer el caso particular de $p = 1$, correspondiente a la distribución exponencial, como posible distribución de los datos. Si realizamos el oportuno contraste, donde representamos gráficamente el logaritmo de la función de supervivencia frente a la variable duración, obtenemos el resultado recogido en la Figura 2.

Por lo tanto, como obtenemos una línea más o menos recta que pasa por el origen, parece que la distribución exponencial es una distribución adecuada para los datos.

A continuación, relajamos el supuesto de población homogénea e introducimos las variables explicativas señaladas anteriormente. Así, analizaremos el efecto que tiene cada una de ellas sobre el tiempo de supervivencia. Dado que la distribución exponencial es un caso particular de la distribución de Weibull y, como además, esta última es lo suficientemente flexible como para englobar distintos tipos de funciones de riesgo (crecientes, decrecientes o constantes en función del valor que tome el parámetro p), parece sensato comenzar ajustando un modelo de regresión Weibull.

Figura 2: Contraste gráfico para una distribución exponencial



El efecto de las variables X sobre la función de supervivencia y riesgo puede incluirse a través del parámetro de escala λ. Para ello, utilizaremos la forma funcional habitual

$$\lambda = e^{-x^T\beta}, \tag{4}$$

donde x^T es el vector (1×10) de valores que toman los regresores, incluyendo la constante, para cada individuo, y β el vector (10×1) de coeficientes asociados a cada regresor. Por tanto, la función de supervivencia quedará especificada como,

$$S(t,x) = \exp[-(e^{-x^T\beta}t)^p], \qquad p > 0, \quad t > 0,$$

y la función de riesgo como

$$\lambda(t,x) = e^{-x^T\beta}p(e^{-x^T\beta}t)^{p-1}, \qquad p > 0, \quad t > 0.$$

Para la especificación (4), este modelo puede reescribirse en términos log-lineales; es decir,

$$\ln(T) = X\beta + \sigma\epsilon, \qquad \text{donde} \qquad \sigma = p^{-1},$$

y donde ϵ tiene una distribución valor extremo estándar. La estimación del modelo se realiza maximizando la función de verosimilitud (1). Los resultados de la estimación se muestran en la Tabla 2

Tabla 2: Estimación del modelo de regresión Weibull

VARIABLE	COEFICIENTE	DESV.TIPICA	T-RATIO	P-VALOR
Constante	1.6864	0.4265	3.954	0.00007
Sexo	0.0585	0.1285	0.455	0.64913
Periodo	0.2024	0.1029	1.967	0.04915
Enfer1	0.1881	0.2455	0.766	0.44364
Enfer2	-0.0247	0.3057	-0.081	0.93558
Sexual	-0.1829	0.2372	-0.771	0.44070
Drogas	-0.0392	0.2031	-0.193	0.84711
Sanguínea	-0.0114	0.2735	-0.042	0.96671
Madre-hijo	0.4005	0.4429	0.904	0.36593
Edad	-0.0172	0.0065	-2.621	0.00876
σ	0.9899	0.0366	26.99	0.00000

Analizando los resultados de la Tabla 2 podemos apreciar una estimación del parámetro σ igual a 0.9899, prácticamente 1, y además, significativo. Esto nos está indicando que la distribución de la duración puede ser una exponencial (como ya anticipábamos). Para contrastar esta hipótesis podemos construir el estadístico correspondiente al contraste de la razón de verosimilitudes. Es decir; realizamos el siguiente contraste dentro de la clase de modelos de regresión Weibull: $H_o: \ \sigma = 1$ (distribución exponencial) frente a $H_a: \ \sigma \neq 1$ (distribución no exponencial).

Ajustamos los modelos bajo la hipótesis nula y bajo la hipótesis alternativa y calculamos el máximo del logaritmo de la función de verosimilitud para cada caso. Obtenemos unos valores de -713.29, en el modelo exponencial, y -713.25 en el modelo no exponencial.

Si construimos el estadístico tenemos que,

$$\Lambda = -2\{\ln[L(\tilde{\beta}, \sigma = 1)] - \ln[L(\hat{\beta}, \hat{\sigma})]\} \xrightarrow{\text{d}} \chi^2_1, \tag{5}$$

donde $(\tilde{\beta}, \sigma = 1)$ son las estimaciones del modelo restringido, en nuestro caso el modelo exponencial, y $(\hat{\beta}, \hat{\sigma})$ son las del modelo general, es decir, del modelo Weibull. Por tanto no encontramos evidencia estadística contraria a la especificación de un modelo de regresión exponencial.

Como consecuencia del contraste, parece razonable ajustar un modelo de distribución exponencial a nuestros datos. Estimamos de nuevo por máxima verosimilitud y obtenemos los resultados recogidos en la Tabla 3.

Si comparamos los resultados de las Tablas 2 y 3, vemos que apenas varían, lo que refuerza la idea de la distribución exponencial.

Podríamos pensar en llevar a cabo un contraste de significación conjunto de todas las variables del modelo, excepto la constante, para estudiar la contribución conjunta de todas las variables sobre el ajuste del modelo. En el caso de que rechazáramos la no significatividad conjunta del modelo, ésta no sería una condición suficiente para considerar al modelo especificado como válido, habríamos de contrastarla con algún contraste de diagnóstico basado en los residuos, lo que realizaremos posteriormente.

Tabla 3: Estimación del modelo de regresión exponencial

VARIABLE	COEFICIENTE	DESV.TIPICA	T-RATIO	P-VALOR
Constante	1.6844	0.4306	3.912	0.00009
Sexo	0.0577	0.1298	0.445	0.65700
Periodo	0.2049	0.1035	1.980	0.04770
Enfer1	0.1873	0.2479	0.755	0.45000
Enfer2	-0.0257	0.3087	-0.083	0.93400
Sexual	-0.1835	0.2396	-0.766	0.44400
Drogas	-0.0397	0.2052	-0.193	0.84700
Sanguínea	-0.0109	0.2763	-0.040	0.96800
Madre-hijo	0.3982	0.4472	0.890	0.37300
Edad	-0.0172	0.0066	-2.607	0.00913
σ	1	-	-	-

Pasamos a realizar el contraste utilizando el estadístico formado por la razón de verosimilitudes. Ajustamos el modelo restrictivo en el que sólo tenemos como variable regresora la constante y obtenemos un valor máximo del logaritmo de la función de verosimilitud de -724.87. Para el modelo menos restrictivo, donde incluimos todas las variables regresoras obtenemos un valor de -713.29. Si calculamos el valor del estadístico para este contraste, que en este caso se distribuye como una χ^2 con 9 grados de libertad, obtenemos un valor de 23.16, superior incluso al cuantil que deja una probabilidad del 1% a su derecha (21.7). Por tanto, podemos concluir que, aún con un nivel de significación del 1%, rechazamos que las variables regresoras en conjunto no contribuyen a la explicación del modelo.

En cuanto al efecto de cada variable, tenemos que el tiempo de supervivencia del individuo se verá afectado por el periodo en el que se le diagnosticó el SIDA, si el diagnóstico del individuo es posterior a 1987, influirá positivamente en su duración. Por tanto, parece que el uso del fármaco zidovudine, más conocido como AZT, alarga el tiempo de supervivencia y reduce el riesgo, obteniendo unos tiempos de supervivencia, a partir del diagnóstico, superiores.

En cuanto a la variable edad, también parece ser relevante para explicar el tiempo de supervivencia del individuo. A mayor edad, menor será el tiempo de supervivencia y, por tanto, mayor el riesgo.

Una vez tenido en cuenta el efecto de estas variables sobre el tiempo de supervivencia, parece que variables como sexo, el tipo de enfermedad con la que se le diagnostica el SIDA o la categoría de transmisión a la que pertenece no influyen significativamente sobre el tiempo que sobrevive el enfermo.

Además, para los modelos de regresión exponencial y para la especificación $\lambda(x,\beta) = \exp(x^T\beta)$, es posible interpretar los coeficientes en términos de la duración media. La media de una variable aleatoria con distribución exponencial y función de densidad $f(t) = \lambda e^{-\lambda t}$, es $1/\lambda$. En nuestro modelo habíamos especificado $\lambda(x,\beta) = e^{-x^T\beta}$, entonces $1/\lambda = e^{x^T\beta}$. Si tomamos logaritmos tenemos que ln(duración media)$= x^T\beta$, de donde

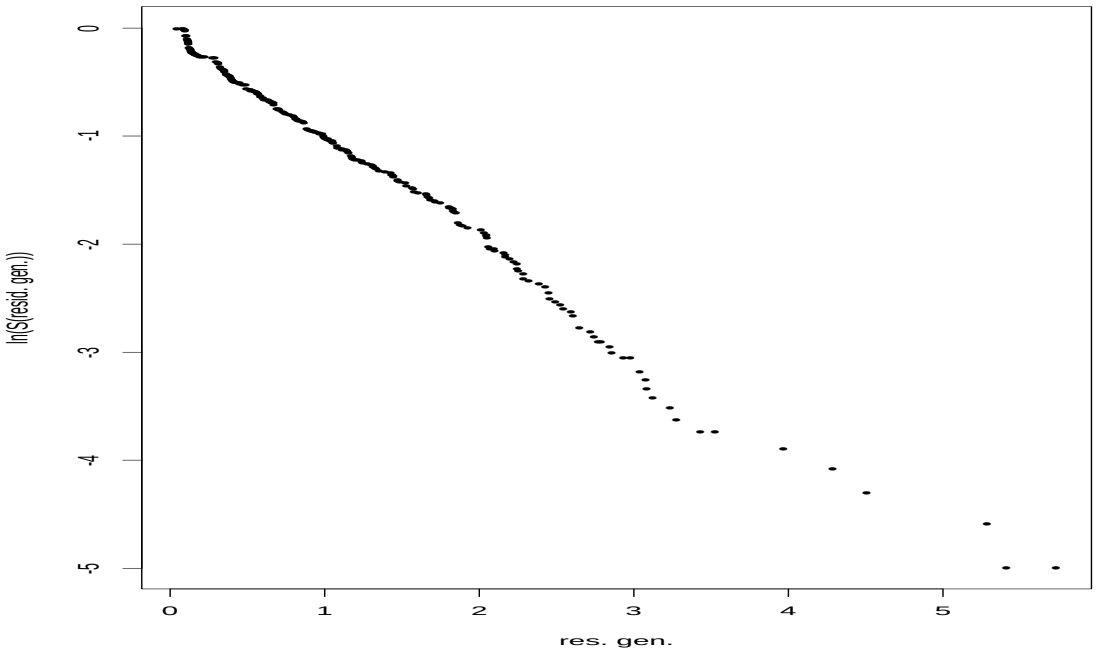
$$\frac{\partial \ln(\text{duración media})}{\partial x} = \beta.$$

Por tanto, β recogerá la variación porcentual de la duración media ante variaciones de la variable regresora x. En el caso de nuestro estudio, esta interpretación sólo tiene sentido para la variable edad (el resto son variables ficticias). Como $\beta_{10} = -0.0172$, este valor nos indica que un aumento de un año en la edad del individuo en el momento de diagnóstico provocará un descenso del 1.7% en el tiempo de supervivencia medio.

Para que los contrastes y conclusiones sobre las estimaciones tengan validez, tenemos que asegurarnos de que el modelo de regresión exponencial se ajusta a nuestros datos realizando un contraste de diagnóstico. Un contraste de diagnóstico sencillo y frecuentemente utilizado es el basado en los residuos generalizados ($\hat{\Lambda}(y_i, x_i)$) o las estimaciones de la función de riesgo integrado[10].

Comenzaremos con un contraste gráfico, donde representamos el logaritmo de la función de supervivencia estimada para los residuos generalizados, mediante el método de Kaplan-Meier, frente a los residuos generalizados. Bajo la hipótesis nula de especificación correcta, debemos obtener una línea recta que pase por el origen, y de pendiente menos uno.

Figura 3: Contraste de diagnóstico



De la Figura 3 podemos suponer que la especificación es aproximadamente la correcta.

Para corroborar el contraste gráfico, y basandonos en la misma idea, realizamos un contraste más formal utilizando el estadístico propuesto por Kiefer (1988)

$$\frac{[\sum_{i=1}^{n} \hat{\Lambda}(y_i, x_i)^2] - 2n}{\sqrt{20n}},$$

[10]Los residuos generalizados $\hat{\Lambda}(y_i, x_i)$ tienen una distribución exponencial estandar bajo la hipótesis nula de especificación correcta del modelo.

con distribución asintótica $N(0,1)$, bajo la hipótesis de correcta especificación del modelo. Antes de computar el estadístico tenemos que ajustar las observaciones censuradas, sumándoles el valor medio[11], en este caso un 1. El valor del estadístico es de 0.13. Por tanto, no encontramos evidencia contraria a la especificación de un modelo de regresión exponencial.

3.3 Estimación mediante un MLG

A continuación volvemos a estimar el modelo de regresión exponencial, pero reescrito como un modelo lineal generalizado, según la propuesta de Aitkin y Clayton (1980).

Construimos el logaritmo de la función de verosimilitud (2) para este caso concreto, obtenemos:

$$\ln L = \sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i) + \sum_{i=1}^{n} \delta_i \ln(y_i^{-1}), \tag{6}$$

donde $\mu_i = \lambda_0 y_i e^{x_i^T \beta}$.

El MLG auxiliar, con una función de verosimilitud proporcional a la anterior, es un modelo donde la variable respuesta δ_i tiene una distribución de Poisson y la relación entre la media de ésta y el predictor lineal se establece, a través de una función de enlace logarítmica, del siguiente modo[12]:

$$\ln \mu_i = \ln y_i + x_i^T \beta$$

La estimación de este MLG puede realizarse utilizando el algoritmo IRLS descrito en la Sección 2. Para proceder con este método necesitamos construir la variable aleatoria Z^*,

$$\ln \delta = \underbrace{\ln \mu + (\delta - \mu) \left.\frac{d(\ln \mu)}{d\mu}\right|_{\mu=\hat{\mu}}}_{z^*} + \ldots .$$

En el modelo de estudio,

$$z^* = \ln y + x^T \beta + (\delta - y e^{x^T \beta}) \left[\frac{1}{y e^{x^T \beta}}\right].$$

Y las ponderaciones a utilizar en cada iteración son:

$$W^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 \Bigg|_{\mu=\hat{\mu}} V(\hat{\mu}) = \frac{1}{y e^{x^T \beta}},$$

donde $V(\hat{\mu})$ es la función de varianza de δ evaluada en $\hat{\mu}$.

[11] Para más detalles sobre este contraste consultar Kiefer (1988).

[12] En un principio el término "offset" viene dado por $\ln(\lambda_0 y_i)$ en lugar del valor $\ln y_i$, especificado en la expresión del modelo. El elemento $\ln \lambda_0$ ha sido introducido en el coeficiente relativo al término constante del predictor lineal.

Iniciamos el proceso tomando como valores iniciales los propios datos, es decir, $\mu_0 = \delta$, construimos $\eta_0 = \ln(\mu_0)$ y, a través de estos valores[13], construimos el vector Z_0^* y la matriz de ponderaciones W_0 para la primera iteración. Continuamos, de forma iterativa, regresando la variable Z^* por mínimos cuadrados sobre los regresores en X y sobre el "offset" $\ln(y)$ utilizando las ponderaciones recogidas en W, hasta alcanzar la convergencia. Como es lógico, obtenemos las mismas estimaciones que las obtenidas en la Tabla 3.

4 Análisis de duración mediante un MLG semiparamétrico

Si analizamos el modelo ajustado en la sección previa nos encontramos un modelo donde el efecto de las variables explicativas es introducido de una forma paramétrica. Es decir, estamos imponiendo una determinada relación (lineal) entre éstas y el logaritmo de la función de riesgo, en el caso de expresar el modelo como caso particular de los modelos de riesgo proporcional, o con el logaritmo de la duración, en el caso de especificar un modelo log-lineal. En algunas situaciones podríamos considerar que la relación paramétrica especificada para alguna de las variables explicativas es muy restrictiva, pudiendo resultar más adecuado introducir el efecto de esta variable de una forma no paramétrica. Así, tendríamos una especificación semiparamétrica, con una parte en la que se recogen variables relacionadas de una forma lineal con la variable a explicar, y otra parte, en la que no se especifique una particular dependencia paramétrica sobre la variable a analizar. Es decir, permitiríamos que los datos reflejaran esta relación mediante una curva de suavizado no paramétrica. Con esta generalización o extensión ampliamos de forma considerable el campo de aplicación de la metodología anterior. Esta extensión nos permite modelizar situaciones en las que no conocemos la forma funcional del efecto de una variable explicativa sobre la variable a explicar, o situaciones en las que suponer una dependencia lineal, u otra cualquiera, entre alguna de las variables explicativas y la variable a analizar sea un supuesto bastante fuerte, o incluso carezca de sentido.

Por tanto, como se puede apreciar la propuesta que vamos a presentar a continuación podría aplicarse en un importante número de situaciones. Un ejemplo ilustrativo del tipo de situaciones que podrían estimarse bajo esta propuesta se recoge en la Sección 3.

En el modelo de la Sección 3 la variable explicativa periodo intenta recoger el efecto de la introducción, a mediados del año 1987, del fármaco zidovudine. Esta variable está construida como una variable ficticia que toma dos valores: valor 1 indicándonos que el diagnóstico del individuo se ha producido con posterioridad a 1987, y valor 0 en caso contrario. Resulta bastante restrictivo dividir el efecto periodo de diagnóstico en dos grupos (antes y después de 1987). Además, parece más lógico o adecuado suponer que el efecto no va ser tan brusco como queda especificado por esa variable ficticia. Por tanto, en esta sección introducimos una componente adicional en el modelo compuesta por una función que depende del periodo de diagnóstico y no especificamos su forma funcional. Así podemos recoger el efecto que tratábamos de recoger, ahora de una forma gradual, además de la evolución completa del

[13]Para evitar el problema de la función de enlace logarítmica aplicada sobre una variable δ que contiene valores cero, hemos sumado una cantidad de 1/6, tal y como se recomienda en Chambers y Hastie (1992).

efecto que tiene el periodo en que se le diagnostica la enfermedad sobre la supervivencia del individuo.

Por tanto, la extensión que estamos proponiendo al trabajo de Aitkin y Clayton (1980) consiste en considerar modelos de duración con la siguiente función de riesgo

$$\lambda(t;x) = \lambda_o(t)\exp(x^T\beta + h(r)), \tag{7}$$

donde $h(r)$ es una función sin especificar con la cual se recoge el efecto de la variable explicativa R. Así, hemos pasado de un predictor lineal paramétrico $\eta = x^T\beta$ a uno semiparamétrico $\eta = x^T\beta + h(r)$.

La estimación, al igual que para el caso paramétrico, se puede realizar utilizando el método de estimación de máxima verosimilitud. Con la diferencia de que en este caso, tenemos que introducir un término adicional en la función de verosimilitud a maximizar. Se trata de un término de penalización asociado a la función h. La idea sobre la necesidad de este término de penalización se basa en que si no establecemos alguna restricción sobre la función h, tenemos a todas aquellas funciones que interpolan los datos como válidas para ajustar el modelo. Para solucionar este problema introducimos una restricción de suavidad para las funciones candidatas a la estimación. Así, tenemos que construir un proceso de estimación que tenga en cuenta tanto la bondad de ajuste del modelo como la suavidad de la función estimadora. Si consideramos como funciones candidatas a todas las funciones pertenecientes al espacio de Sobolev de orden m ($W_2^m[a,b]$). Es decir, todas aquellas funciones cuya derivada m-ésima al cuadrado es integrable en el intervalo $[a,b]$. Y de entre todas éstas, queremos una que se ajuste a los datos y que a su vez sea suave. Por tanto, necesitamos un criterio de elección que considere estas dos características: buen ajuste y suavidad.

La bondad de ajuste dependerá del criterio de optimización elegido. Para el caso de la estimación por máxima verosimilitud, éste estará recogido en la función de verosimilitud.

La medida de suavidad para funciones $h \in (W_2^m[a,b])$ puede estar recogida por $\int_a^b \left[h^{(m)}(r)\right]^2 dr$. En la práctica lo habitual es considerar el caso $m = 2$.

Así, la estimación del modelo puede realizarse mediante la función de verosimilitud penalizada (Good y Gaskins, 1971), la cual considera o tiene en cuenta estas dos características. Para el modelo que estamos considerando, modelo (7), el logaritmo de la función de verosimilitud penalizada tiene la siguiente expresión,

$$\Pi = \sum_{i=1}^{n}\left\{\delta_i\left[\ln\Lambda_0(y_i) + x_i^T\beta + h(r_i)\right] - \Lambda_0(y_i)e^{x_i^T\beta+h(r_i)} + \delta_i\ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)\right\} - \frac{1}{2}\alpha\int_a^b\left[h''(r)\right]^2 dr. \tag{8}$$

El parámetro de suavizado α, refleja la importancia que damos a la suavidad de la función y a la bondad del ajuste del modelo. Para un valor de α grande estamos dando más importancia a la suavidad, penalizando fuertemente las funciones estimadoras con segunda derivada elevada. Para un valor pequeño estamos dando mayor importancia al buen ajuste del modelo.

De forma análoga al tipo de modelo considerado en la Sección 2, la estimación puede realizarse construyendo un MLG auxiliar, en este caso semiparamétrico, con un logaritmo de la función de verosimilitud proporcional a (8).

El MLG semiparamétrico, auxiliar a este modelo de duración concreto, es un modelo log-lineal de Poisson para la variable indicador de censura δ, con una función de enlace logarítmica y el siguiente predictor lineal:

$$\ln \mu_i = \ln \Lambda_0(y_i) + x_i^T \beta + h(r_i)$$

El logaritmo de la función de verosimilitud penalizada de este modelo auxiliar viene dado por:

$$\Pi = \sum_{i=1}^{n} \delta_i \ln \mu_i - \mu_i - \sum_{i=1}^{n} \ln \delta_i! - \frac{1}{2}\alpha \int_a^b \left[h''(r)\right]^2 dr. \tag{9}$$

Si se sustituye μ_i por su valor $e^{\ln \Lambda_0(y_i) + x_i^T \beta + h(r_i)}$, se puede comprobar que esta expresión es proporcional a (8). Por tanto, las estimaciones en una y otra expresión son las mismas.

Antes de pasar a maximizar la expresión (8), señalaremos que se puede demostrar que la solución, para la función h, al problema de maximizar (8) es una función "spline" cúbica natural. De esta forma y utilizando las propiedades de este tipo de funciones podemos reexpresar (8) (y de forma equivalente (9)) como

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T \beta + (Nh)_i \right] - \Lambda_0(y_i) e^{x_i^T \beta + (Nh)_i} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} - \frac{1}{2}\alpha h^T K h, \tag{10}$$

donde ahora h es el vector de valores $h_j = h(r_j)$, para $j = 1, \ldots, d$ donde d indica el número de valores distintos que toma la variable R, la matriz N se conoce como la matriz incidencia y su función consiste en asignar a cada elemento el valor que le corresponde de la variable que hemos introducido de forma no paramétrica y K es una matriz que se construye utilizando ciertas propiedades de las funciones spline cúbicas naturales[14].

Para maximizar respecto a los coeficientes β y a h podemos utilizar el algoritmo de Fisher scoring. La aplicación de este algoritmo, como se puede demostrar (la demostración en detalle puede encontrarse en Orbe, 2000, pag. 144-146), es equivalente a la resolución del siguiente sistema de ecuaciones simultáneas

$$\begin{aligned} X^T W X \beta &= X^T W (Z^* - Nh) \qquad (a) \\ (N^T W N + \alpha K) h &= N^T W (Z^* - X\beta) \qquad (b) \end{aligned} \tag{11}$$

donde el elemento i-ésimo del vector Z^* es:

$$z_i^* = \ln \Lambda_0(y_i) + x_i^T \beta + (Nh)_i + (\delta_i - \mu_i)\frac{1}{\mu_i} \tag{12}$$

donde $\mu_i = \Lambda_0(y_i) e^{x_i^T \beta + (Nh)_i}$, W es una matriz de ponderaciones donde los elementos de la diagonal principal son de la forma[15]

[14] Para más detalles véase, por ejemplo, Green y Silverman (1994), Cap. 2.

[15] Como es lógico μ_i y w_{ii} coinciden, puesto que, por una parte, tenemos una función enlace logarítmica y, por otra parte, en una distribución de Poisson la media y varianza coinciden. Entonces, el paso 2 del algoritmo IRLS nos lleva a la relación $\mu_i = w_{ii}$.

$$w_{ii} = \Lambda_0(y_i)e^{x_i^T\beta+(Nh)_i}, \tag{13}$$

Para obtener las estimaciones de β y h podemos realizar un procedimiento de "backfitting" (Buja, Hastie y Tibshirani, 1989) entre las ecuaciones (11a) y (11b), hasta alcanzar la convergencia. Así, por una parte, si en la ecuación (11a) conocemos h, el vector de coeficientes β se obtiene regresando por mínimos cuadrados ponderados las diferencias $(Z^* - Nh)$ sobre la matriz de variables regresoras X (las variables de la componente paramétrica). Las ponderaciones serán las anteriormente indicadas. Por otra parte, si conocemos β en (11b) podemos obtener h mediante un suavizador spline cúbico natural aplicado a las diferencias $(Z^* - X\beta)$.

Resumiendo, el procedimiento de estimación completo comienza con la construcción de la matriz incidencia N. Posteriormente, iniciamos el proceso iterativo tomando $\hat{\beta} = 0$, calculamos las estimaciones iniciales del vector h regresando por mínimos cuadrados ordinarios el logaritmo de la variable indicadora de censura $\ln\delta$ sobre la matriz de incidencia. Es decir, $\hat{h} = (N^TN)^{-1}N^T(\ln(\delta))$. Con estas dos estimaciones iniciales, construimos la estimación inicial de $\hat{\mu}_i = \Lambda_0(y_i)e^{x_i^T\hat{\beta}+(N\hat{h})_i}$ y, aplicando la función de enlace logarítmica, obtenemos el valor inicial del predictor lineal $\hat{\eta}_i = \ln\Lambda_0(y_i) + x_i^T\hat{\beta} + (N\hat{h})_i$. Utilizando las estimaciones de μ y η, obtenemos el vector Z^* siguiendo la expresión (12) y la matriz de ponderaciones W siguiendo la expresión (13). Una vez obtenidos estos valores iniciales, comenzamos con el procedimiento de backfitting, sustituyendo de forma alternativa las estimaciones de (11a) y (11b) hasta que se produzca la convergencia.

En nuestro caso hemos visto que es adecuado proponer una distribución exponencial para la variable T. Por lo tanto, en la expresión (10) sustituimos las expresiones de las funciones de riesgo y riesgo acumulado por las correspondientes de una variable con distribución exponencial. Así, $\Lambda_0(t) = t$ y $\lambda_0(t)/\Lambda_0(t) = 1/t$. Y para maximizar el sistema (11) previamente debemos de realizar la misma sustitución en las expresiones (12) y (13).

5 Análisis de las estimaciones

Una vez estimado los parámetros del modelo y la función no paramétrica, se nos presenta el problema del análisis de la significatividad o, en general, de realizar inferencia. Dada la componente no paramétrica, podríamos pensar en utilizar contrastes asintóticos (Hastie y Tibshirani, 1990). En lugar de utilizar este tipo de contrastes hemos optado por realizar el estudio de las estimaciones obtenidas mediante técnicas bootstrap. Una de las ventajas que presenta el bootstrap es la posibilidad de analizar las propiedades y realizar inferencia incluso con tamaños de muestra reducidos. Sin embargo, no existe un método bootstrap específico adaptable al modelo propuesto, por lo que procedemos a la elaboración de uno.

Aplicaremos un bootstrap en regresión, ya que disponemos de un conjunto de observaciones no homogéneas, donde la heterogeneidad la tratamos de recoger a través de una serie de variables explicativas utilizando un modelo de regresión. Además, al suponer una distribución concreta para la variable de interés, desarrollaremos un bootstrap paramétrico.

La idea del bootstrap en regresión es la misma que la del bootstrap para modelos homogéneos. Dado que el modelo que estamos considerando parece el adecuado para nuestros

datos, realizamos un bootstrap en regresión basado en el modelo. Este procedimiento consiste en obtener la remuestra bootstrap para la perturbación del modelo y, siguiendo la especificación del modelo, construir la remuestra bootstrap para la variable respuesta (para más detalles sobre las técnicas bootstrap, véase, por ejemplo, Efron y Tibshirani, 1993, y Davison y Hinkley, 1997).

Por otra parte, dado que en la muestra tenemos observaciones censuradas y esto tiene que reflejarse en las remuestras bootstrap, tenemos que aplicar un bootstrap adecuado para datos censurados. Tenemos dos posibilidades de remuestreo en el caso de muestras con censura. Efron (1981) propone estimar las funciones de distribución Kaplan-Meier (Kaplan y Meier, 1958) para la variable de interés $\hat{F}_n$ y lo mismo para la variable censura $\hat{G}_n$, posteriormente generar con ambas funciones de distribución sendas muestras para la variable de interés $t_1^*, \ldots, t_n^*$ y para la variable censura $c_1^*, \ldots, c_n^*$, y considerar la siguiente remuestra bootstrap,

$$y_i^* = \min(t_i^*, c_i^*), \qquad \delta_i^* = \begin{cases} 1; & \text{si } t_i^* \leq c_i^* \\ 0; & \text{si } t_i^* > c_i^* \end{cases} .$$

La otra posibilidad, presentada por Reid (1981), consiste en tomar una muestra de observaciones independientes e idénticamente distribuidas con la función de distribución, estimada mediante el estimador Kaplan-Meier, de la variable de interés y considerar la correspondiente función de distribución empírica.

Akritas (1986) demuestra que el plan de remuestreo de Efron es mejor que el de Reid. Además, para el caso de censura aleatoria, Efron demuestra que realizar lo anterior es equivalente a remuestrear con reemplazamiento sobre los pares de variable observada e indicador de censura $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.

Hay que señalar que estos dos procedimientos de generación de muestras bootstrap, para muestras con observaciones censuradas, están pensados para el caso de muestras homogéneas; es decir, para situaciones en las que no tenemos variables explicativas que influyen sobre la variable a explicar, y para el caso en que desconocemos las funciones de distribución de la variable de interés y de la variable censura. Sin embargo éste no es nuestro caso. En nuestro problema, estamos suponiendo una distribución para la variable de interés T (distribución exponencial) y no estamos suponiendo distribución alguna para la variable censura C y, además, tenemos variables explicativas en el modelo. Por lo tanto, para solucionar este problema, tenemos que proponer un nuevo procedimiento generador de muestras bootstrap, adecuado a los supuestos del modelo. Hay que señalar que la propuesta de Efron (para el caso de no suponer la distribución de la variable duración y la variable censura), aún seguiría siendo valida para el caso heterogéneo siempre y cuando supongamos que la variable censura siga el mismo modelo de regresión propuesto para la variable duración.

El modelo concreto que estamos considerando es un modelo de regresión exponencial semiparamétrico, es decir, tenemos la siguiente función de densidad para la variable de interés T:

$$f(t; x) = \lambda e^{-\lambda t}; \quad \text{donde} \quad \lambda = e^{-(x^T\beta + h(r))}.$$

Para la variable censura C no estamos suponiendo distribución alguna.

Para este tipo de modelos proponemos el siguiente procedimiento para generar las remuestras bootstrap:

Paso 1: Ajustar el modelo (7) para el caso de una distribución exponencial y obtener los residuos.

Paso 2: Generar las perturbaciones bootstrap $\epsilon_1^*, \ldots, \epsilon_n^*$, con una distribución valor extremo mínimo.

Paso 3: Obtener la muestra bootstrap para la variable de interés basándonos en el modelo

$$\ln T_i^* = x_i^T \hat{\beta} + \hat{h}(r_i) + \epsilon_i^*; \quad \text{para} \quad i = 1, \ldots, n.$$

Paso 4: Obtener la muestra bootstrap para variable censura generando una muestra de n observaciones a partir de la función de distribución G de la variable censura.

Paso 5: Comparando las remuestras bootstrap para la variable de interés (paso 3) y la variable censura (paso 4), obtenemos la variable observada bootstrap Y^*, y la correspondiente variable indicador bootstrap δ^*,

$$y_i^* = \min\{t_i^*, c_i^*\}, \qquad \text{y} \qquad \delta_i^* = \begin{cases} 1; & \text{si } t_i^* \leq c_i^* \\ 0; & \text{si } t_i^* > c_i^* \end{cases} .$$

Paso 6 Estimar el modelo (7) (para el caso exponencial) utilizando la información disponible en la remuestra bootstrap.

Paso 7: Volver al paso 2 y repetir el proceso M veces.

Para obtener las estimaciones del modelo (7), en el paso 1, desarrollamos el proceso de estimación propuesto en la sección anterior para el caso particular de una distribución exponencial para la variable T. En el paso 2 estamos considerando el modelo lineal para el logaritmo de la duración[16]. Por lo tanto, en este paso, obtenemos la remuestra bootstrap de las perturbaciones realizando un bootstrap paramétrico, donde consideramos una distribución valor extremo para las perturbaciones. En el paso 3, como acabamos de comentar, utilizamos la expresión log-lineal de nuestro modelo (7) (ver pie de página 16) para obtener la remuestra bootstrap de la variable de interés. En el paso 4, generamos la variable censura sin considerar ningun supuesto adicional al modelo, que contemple una relación determinada entre las variables explicativas y ésta. La función de distribución G de la variable censura es desconocida y la estimamos utilizando el estimador de Kaplan-Meier, $\hat{G}_n$, adecuado para esta variable. En el paso 6, y como en el paso 1, utilizamos el procedimiento de estimación descrito en la Sección 4. Por último, indicaremos que el número de remuestras bootstrap a considerar depende del objetivo del estudio, si únicamente deseamos calcular las desviaciones típicas de las estimaciones obtenidas, un valor de $M = 200$ puede ser suficiente para obtener unos valores fiables. En cambio, si nuestro objetivo es más ambicioso, y deseamos construir intervalos de confianza, tenemos que considerar un número sensiblemente superior (al menos $M = 1000$), para tener una buena estimación de los percentiles en las colas de la distribución.

[16] Como suponemos una distribución exponencial de parámetro $\lambda = e^{-(x^T\beta + h(r))}$ para la variable duración, al tomar la transformación logarítmica podemos reescribir el modelo en términos log-lineales como $\ln T = X\beta + h(r) + \epsilon$ donde, entonces, ϵ tiene una distribución valor extremo mínimo.

6 Resultados y conclusiones

Como ilustración de las dos secciones anteriores aplicamos la metodología descrita al conjunto de datos presentados en la Sección 3. Así, la motivación de la extensión del modelo paramétrico, ajustado en la Sección 3, a uno semiparamétrico, tiene su fundamento en el supuesto más razonable de un efecto real, de la introducción del fármaco AZT, más suave o más gradual que el especificado en la Sección 3, utilizando una variable ficticia. Por lo tanto, ahora consideramos un modelo semiparamétrico, concretamente ajustamos el modelo (7) para el caso particular de una variable T con distribución exponencial. El efecto de las variables explicativas quedará dividido en dos términos. El paramétrico, donde recogemos el efecto de todas las variables explicativas excepto la variable periodo de diagnóstico[17] cuyo efecto sera recogido de una forma no paramétrica a través de una función h.

Los resultados de esta estimación y del posterior análisis, de las estimaciones obtenidas mediante técnicas bootstrap, son presentados en las Tablas 4, 5 y Figura 4.

Tabla 4: Estimaciones de los coeficientes β de la componente paramétrica

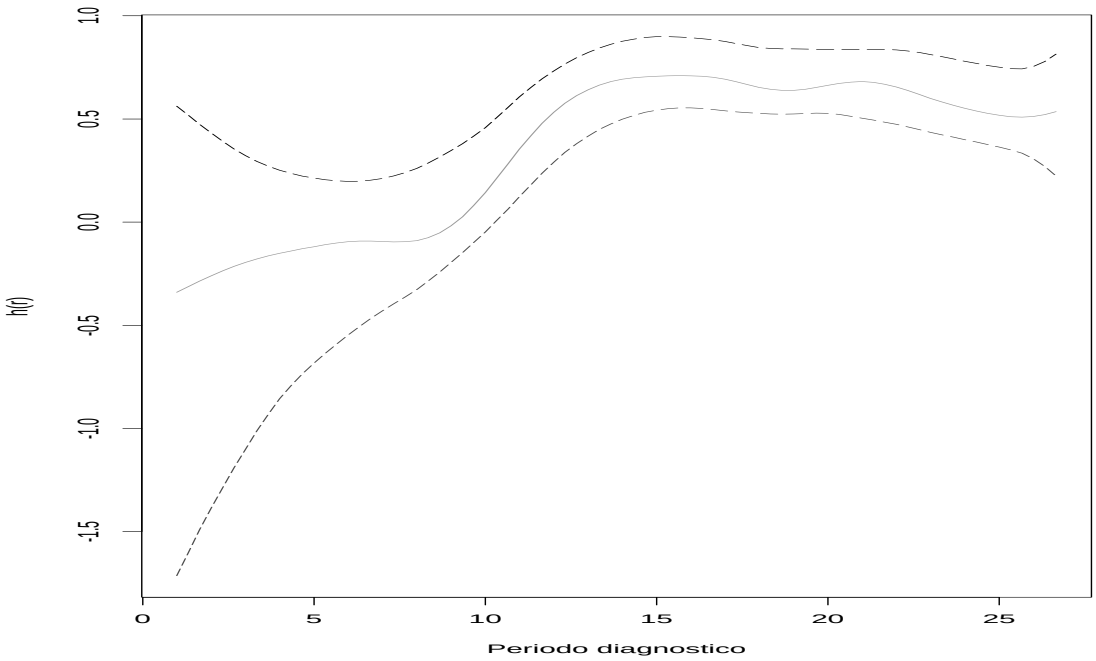
VARIABLE	COEFICIENTE	DESV.TIPICA
Constante	1.37063	0.40330
Sexo	0.02259	0.13103
Enfer1	0.19599	0.25522
Enfer2	0.13172	0.31948
Sexual	-0.23432	0.24722
Drogas	-0.10928	0.20709
Sanguínea	-0.00008	0.28563
Madre-hijo	0.17904	0.47801
Edad	-0.01870	0.00643

Tabla 5: Intervalos de confianza bootstrap (95%) para las estimaciones de los coeficientes β

VARIABLE	LIM. INF.	LIM. SUP.
Constante	0.5391	2.1134
Sexo	-0.2176	0.2659
Enfer1	-0.2796	0.7088
Enfer2	-0.4584	0.7571
Sexual	-0.6794	0.2768
Drogas	-0.4636	0.3413
Sanguínea	-0.5258	0.5947
Madre-hijo	-0.6230	1.1253
Edad	-0.0308	-0.0056

[17]Ahora, a diferencia de la Sección 3, la variable periodo no va a ser una variable ficticia. En su lugar vamos a crear una nueva variable periodo de diagnóstico que toma valor 1 para los individuos diagnosticados de SIDA en el primer trimestre de la muestra, valor 2, para los diagnosticados en el segundo trimestre, y así sucesivamente, hasta el último trimestre de diagnóstico existente en la muestra.

Figura 4: Estimación e intervalo de confianza bootstrap (95%) para la componente no paramétrica



Indicaremos que el número de remuestras bootstrap considerado para este análisis es de $M = 1999$.

La Tabla 4 muestra la estimación de los coeficientes β para aquellas variables introducidas en la componente paramétrica del modelo junto a la estimación bootstrap de sus desviaciones típicas. La Tabla 5 presenta los intervalos de confianza bootstrap percentil BC (al 95%) para estos coeficientes β. La Figura 4 nos presenta la estimación de la componente no paramétrica, la función $h(r)$, así como, las bandas de confianza bootstrap percentil al 95% (para una detallada descripción sobre intervalos de confianza bootstrap ver, por ejemplo, Efron, 1987 y Efron y Tibshirani, 1986).

Antes de pasar a interpretar los resultados obtenidos tenemos que señalar que las estimaciones presentadas en las tablas y figura mencionadas anteriormente indican el efecto de esas variables sobre el logaritmo de la duración. El efecto sobre la función de riesgo va a ser el mismo pero de signo contrario al presentado en las tablas y figura. Una vez aclarada esta cuestión pasamos a reseñar los resultados más relevantes.

En cuanto a las variables introducidas en la componente paramétrica del modelo, indicaremos que únicamente la variable edad resulta significativa para explicar el tiempo de supervivencia del enfermo. A mayor edad en el momento del diagnóstico tenemos un tiempo de supervivencia menor para el enfermo. El resto de las variables de la componente paramétrica resultan no significativas para explicar la supervivencia. Estos mismos resultados se han obtenido en otros trabajos como se recoge en la síntesis de resultados, obtenidos por diferentes autores aplicando diferentes metodologías, presentada en Brookmeyer y Gail (1993).

En cuanto a la componente no paramétrica, propuesta para flexibilizar la más restrictiva

aproximación realizada en la Sección 3 (donde se dividía el periodo de estudio en dos partes mediante una variable ficticia), podemos apreciar además del efecto de la introducción del fármaco AZT, la evolución del efecto del periodo de diagnóstico sobre el tiempo de supervivencia. Así, podemos observar una tendencia ligeramente creciente, mayores tiempos de supervivencia, a medida que nos desplazamos de los primeros periodos de diagnóstico. Esta suave tendencia creciente puede venir provocada por el cada vez mayor conocimiento de la enfermedad con el paso del tiempo, lo cual puede originar diagnósticos cada vez más precoces, aumentando así el tiempo de supervivencia desde el momento del diagnóstico. Posteriormente, observamos una fuerte aceleración, en este efecto positivo, sobre la supervivencia, para finalmente mantenerse en niveles máximos. Aqui habría que recordar que la introducción del AZT se produce a mediados de 1987 (alrededor del trimestre 13). Por lo tanto, la Figura 4 parece mostrarnos un efecto beneficioso de la introducción del fármaco, provocando una importante mejora en el tiempo de supervivencia del enfermo. Como se puede apreciar en la figura, la aceleración de este efecto positivo se produce varios trimestres antes de la introducción del fármaco, lo cual resulta bastante lógico, ya que individuos diagnósticados de SIDA, antes de la introducción del fármaco, también van a recibir el fármaco (aunque no desde un principio) y por tanto, también se benefician de los resultados positivos de éste. Señalaremos que este efecto positivo del AZT también se obtiene en el modelo de la Sección 3 y en otros trabajos como se señala en Brookmeyer y Gail (1993). Entre ellos podemos citar, por ejemplo, Lemp y otros (1990) y Moore y otros (1991). Sin embargo, tenemos que añadir que con la especificación semiparamétrica que proponemos en la Sección 4 somos capaces de capturar el efecto del AZT de una forma gradual y más flexible, cosa que no podemos hacer bajo una especificación con variable ficticia, puesto que esta especificación esta considerando un efecto repentino o brusco. Además, la especificación semiparamétrica nos permite, aparte del efecto de la introducción del AZT, analizar la evolución total del efecto periodo de diagnóstico sobre la supervivencia.

Antes de finalizar, indicaremos que la principal motivación del trabajo presentado ha sido la propuesta de extensión del trabajo de Aitkin y Clayton (1980). Por tanto, el análisis empírico llevado a cabo, prentende, básicamente, ilustrar la metodología propuesta. Aún así, se han extraido una serie de resultados interesantes que, además, nos pueden ayudar a entender mejor la relevancia de la propuesta que estamos realizando. La extensión propuesta amplía el campo de aplicación de la metodología de esos autores, permitiendo considerar aquellas situaciones donde la forma funcional del efecto de alguna de las variables explicativas sobre la variable de interés es desconocida o situaciones en las que la especificación de una determinada forma funcional resulta un supuesto bastante restrictivo o carece de sentido. Para finalizar, señalaremos que la inferencia del modelo se ha realizado mediante técnicas bootstrap, para lo cual hemos propuesto un procedimiento de generación de remuestras bootstrap adecuado a las características del modelo.

7 Referencias

Aitkin M. & Clayton D. (1980). The Fitting of Exponential, Weibull and Extreme Value Distributions to Complex Censored Survival Data using GLIM. *Applied Statistic*, **29**, 156-163.

Akritas M.G. (1986). Bootstrapping the Kaplan-Meier Estimator. *Journal of the American Statistical Association*, **81**, 1032-1038.

Brookmeyer R. & Gail M. H. (1993). *AIDS Epidemiology a Quantitative Approach*. Oxford University Press: Oxford.

Buja A., Hastie T. J. & Tibshirani R. J. (1989). Linear Smoothers and Additive Models (with Discussion). *Annals of Statistic*, **17**, 453-555.

Chambers J. M. & Hastie T. J. (1992). *Statistical Models in S*. Wadsworth and Brooks: Pacific Grove, California.

Cox D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society-Series B*, **34**, 187-220.

Cox D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Davison A. C. & Hinkley D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press: Cambridge.

Efron B. (1981). Censored Data and Bootstrap. *Journal of the American Statistical Association*, **76**, 312-319.

Efron B. & Tibshirani R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, **1**, 54-77.

Efron B. & Tibshirani R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall: New York.

Efron B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, **82**, 171-200.

Fahrmeir L. & Tutz G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag: New York.

Good I. J. & Gaskins R. A. (1971). Non-parametric Roughness Penalties for Probability Densities. *Biometrika*, **58**, 255-277.

Greenwood M. (1926). The Natural Duration of Cancer. *Reports on Public Health and Medical Subjects*, **33**, 1-26.

Green P. J. & Silverman B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall: London.

Hastie T.J. & Tibshirani R.J. (1990). *Generalized Additive Models*. Chapman and Hall: London.

Kalbfleisch J. D. & Prentice R. L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons: New York.

Kaplan E. L. & Meier P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457-481.

Kiefer N. M. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, **26**, 646-679.

Lawless J. F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons: New York.

Lemp G. P., Payne S. F. & Neal D. (1990). Survival Trends for Patients with AIDS. *Journal of the American Medical Association*, **263**, 402-406.

McCullagh P. & Nelder J. A. (1983). *Generalized Linear Models*. Chapman and Hall: London.

Moore R. D., Hidalgo J., Sugland B. W. & Chaisson R. E. (1991). Zidovudine and the Natural History af the Acquired Inmunodeficiency Syndrome. *New England Journal of Medicine*, **263**, 1412-1416.

Nelder J. A. & Wedderburn R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society-Series A*, **135**, 370-384.

Orbe J. (2000). *Un Modelo de Regresión Parcial Censurado para Análisis de Supervivencia*. Tesis Doctoral, Universidad del País Vasco, Bilbao.

Reid N. (1981). Estimating the Median Survival Time. *Biometrika*, **68**, 601-608.

Whitehead J. (1980). Fitting Cox's Regression Model to Survival Data using GLIM. *Applied Statistics*, **29**, 268-275.